INDENG242 Final Project

Predict the winning team in the League of Legends World Championship 2023

Xilin Tian, Xinyu Hou, Yunqi Liang, Jiayi Fang

Master of Analytics

IEOR
University of California, Berkeley
Fall 2023

Contents

1 Abstract **2**

2 Introduction **2**

3 Data Collection **3**

4 Analytical Progress **3**
4.1 Logistic Regression 3
4.2 CART 5
4.3 Random Forest 5
4.4 Boosting 6
4.5 K-mean 6

5 Bootstrapping **7**

6 Conclusion **8**

7 Discussion **8**

8 Reference **9**

1 Abstract

As a MOBA game with players all over the world, League of Legends is a prominent achievement in the gaming industry. And their most important international tournament, Worlds, has attracted the attention of a very large number of players and the participation of many professional players. This report is based on the average match data of the participating teams this year to predict the Worlds 2023 World Champion. By applying different models and adding new variables, we achieved an accuracy of about 70% and one of the models, XGBoost, accurately predicted the realistic winner: T1.

2 Introduction

League of Legends is a team-based strategy game where two teams of five powerful champions face off to destroy the other's base. Choose from over 140 champions to make epic plays, secure kills, and take down towers as you battle your way to victory.[1] In the game, two teams of five players battle in player-versus-player combat, each team occupying and defending their half of the map. Each of the ten players controls a character, known as a "champion", with unique abilities and differing styles of play. During a match, champions become more powerful by collecting experience points, earning gold, and purchasing items to defeat the opposing team. In League's main mode, Summoner's Rift, a team wins by pushing through to the enemy base and destroying their "Nexus", a large structure located within.[2]

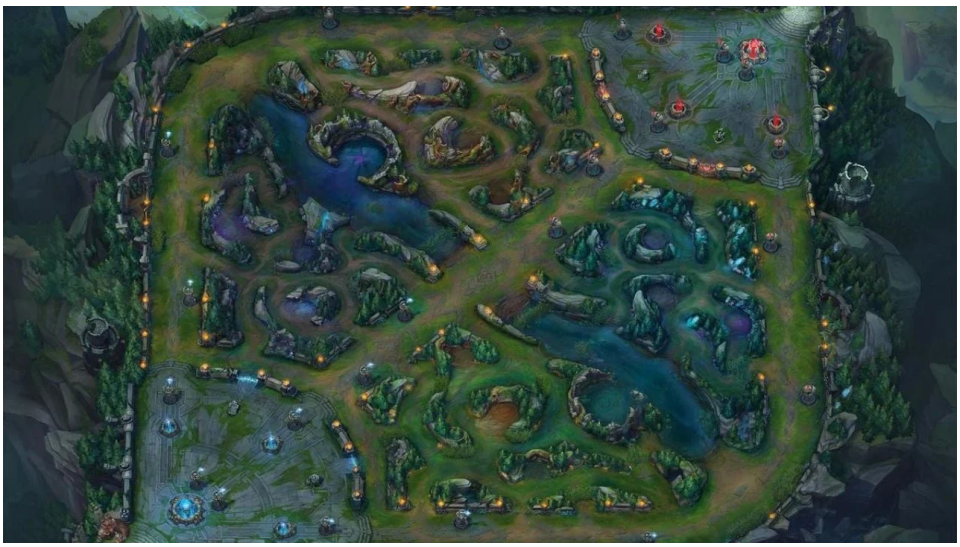

Figure 1: League of Legends 5v5 map

The 2023 Season World Championship (Worlds 2023) is the conclusion of the 2023 League of Legends esports season. The tournament will be held in South Korea. [3] When a new match begins, teams choose either the blue side or the red side based on a roll of the dice or other fair method, with the blue team having priority in disabling heroes and choosing heroes, and the red team getting to choose the role of counter based on the blue side's choice. It's easier for the blue team to approach the Rift Vanguard and Baron than it is for the red team. At the same time, however, Red does have easier access to the Dragon Pit. Considering this slight advantage and the stats throughout the years in League of Legends, Blue has a higher win rate than Red. [4] So based on the system where the loser of the previous game chooses the red and blue side for the next game, let's assume that each team prioritizes the blue side. In the final match of Worlds 2023, T1 3-0 edged out WBG to win the Championship. This is T1's fourth Champion, as well as the fourth for Faker.

3 Data Collection

The records that we collected for each team for the entire year from oracle[5] and wins and losses for each game of the Worlds from League of Legends Wiki[6]. From 2019 to 2023 for total 5 years, we have 368 matches; and for each team, total 21 attributes for each year.

In order to join the datasets of two opposing teams together, We first calculate the Win Rate to replace with the number of Winning Game, Losing Game and Total Game by dividing Total Game from Winning Game:

$$\mathbf{WR} = \frac{\mathbf{\#\ of\ Winning\ Games}}{\mathbf{\#\ of\ Total\ Games}}$$

Then we subtract the data of the red team from the blue team for each game to get the difference between the two teams as the INDEPENDENT variable of the dataset; and whether the blue team wins or not as the DEPENDENT variable of the dataset. Thus, we obtain a new dataset that will be based on the predicted the winning rate of the blue team for further analysis.

WR	winning percentage
KD	Kill-to-Death Ratio
CKPM	Average combined kills per minute (team kills + opponent kills)
GPR	Gold percent rating
GSPD	Average gold spent percentage difference
EGR	Early-Game Rating
MLR	Mid/Late Rating
FB	First Blood rate
FT	First tower rate
F3T	First-to three-towers rate
HLD	Rift Herald control rate
FD	First dragon rate
DRG	Dragon control rate
ELD	Elder dragon control rate
BN	Baron control rate
LNE	Lane Control
JNG	Jungle Control
WPM	Average wards placed per minute
CWPM	Control wards purchased per minute

Table1: INDEPENDENT Variables

4 Analytical Progress

Our project aims to predict the winner of Worlds 2023 using logistic regression, CART, random forest, and XGBoost models. To refine our model, we employ K-mean, and use Bootstrapping to select the best model. Our prediction progress focuses on predicting the knockout stage, which is the game between the top eight teams.

For each matches, we followed the schedule from Worlds, subtracting the data between the two teams from each other to get the difference between each and the other when they were on the blue side. The combined data is substituted into the model to derive a win rate for comparison, and the winner between each two teams moves on to the next round for the competition.

4.1 Logistic Regression

Since win or lose is a binary categorization problem, as ONE of the most efficient algorithms, we first use Logistic Regression under the package of statsmodel for the fitted models. After splittig into Training and Testing Dataset, the OLS Summary for original dataset is showing below(Figure 2). This

Summary gives us a generalized idea about the relevance of each variable to the ability to win the game. For example, we can see that the coefficient for WR and LNE are positive and its corresponding p-value is samll, which means the difference in Winning Rate and LaNE control between the two teams plays a very important role in whether the game is won or lost. Also, considering the existence of multicollinearity, we removed the high VIF and high p-value variables and applied in the testing set to observe the accuracy, TPR, and FPR of each model, and ultimately, because the Logistic Regression model with the high p-value removed possessed the highest accuracy (64.8%), we used this model to predict the Worlds 2023 Top-eight matches.

Dep. Variable:	IsWin	No. Observations:	294
Model:	Logit	Df Residuals:	274
Method:	MLE	Df Model:	19
Date:	Mon, 20 Nov 2023	Pseudo R-squ.:	0.1424
Time:	00:08:14	Log-Likelihood:	-174.61
converged:	True	LL-Null:	-203.62
Covariance Type:	nonrobust	LLR p-value:	7.973e-06

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.0932	0.131	0.712	0.476	-0.163	0.350
WR	14.2685	5.686	2.509	0.012	3.124	25.413
KD	-3.9145	1.695	-2.310	0.021	-7.236	-0.593
CKPM	-3.0295	1.631	-1.857	0.063	-6.227	0.168
GPR	1.6363	1.026	1.595	0.111	-0.375	3.648
GSPD	-24.3825	13.298	-1.834	0.067	-50.446	1.681
EGR	-0.0471	0.067	-0.701	0.483	-0.179	0.085
MLR	0.0087	0.035	0.251	0.802	-0.059	0.077
FB	2.6093	1.949	1.339	0.181	-1.211	6.430
FT	2.7085	1.870	1.449	0.147	-0.956	6.373
F3T	-4.6599	2.240	-2.080	0.038	-9.051	-0.269
HLD	3.0294	1.586	1.910	0.056	-0.080	6.139
FD	-2.8622	2.086	-1.372	0.170	-6.950	1.226
DRG	-1.9032	4.414	-0.431	0.666	-10.554	6.747
ELD	-0.2659	0.524	-0.507	0.612	-1.294	0.762
BN	1.4136	2.829	0.500	0.617	-4.132	6.959
LNE	50.5712	30.048	1.683	0.092	-8.322	109.465
JNG	8.7887	11.313	0.777	0.437	-13.384	30.962
WPM	0.4198	0.694	0.605	0.545	-0.940	1.780
CWPM	1.5248	1.262	1.208	0.227	-0.949	3.998

Figure 2: Logistic Regression Summary for All Variables

The following Tree Diagram(Figure 3) shows the win rate for that team on the blue side in each game between teams as defined by the tournament system starting with the Top 8. The winner of the final logistic regression model prediction was Gen. G.

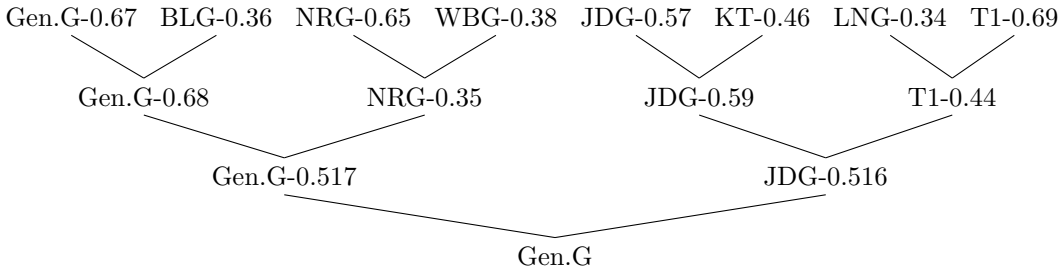

Figure 3: Tree Diagram for Predicting Winner by Logistic Regression Model

4.2 CART

As an advantage of the same transparent, easy-to-understand approach, CART served as the model for our second experiment. In the selection of features, we use the importance from the sklearn package to drop the three least important variables. After pruning the CART model by using the optimal ccp_alpha obtained by GridSearchCV, we obtained a model with 58% accuracy. The winner of the final CART model prediction shown below(Figure 4) was NRG.

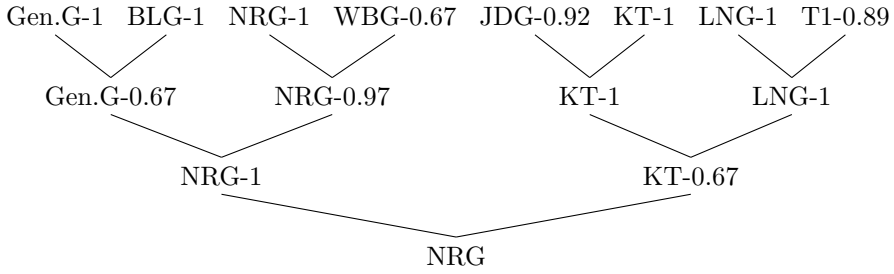

Figure 4: Tree Diagram for Predicting Winner by CART Model

4.3 Random Forest

Previously we applied the CART model, and Random Forest, which combines the outputs of multiple decision trees to produce a single result, will be the model we'll be using soon. Although Random forest does not require feature selection, we made correlation icons for each variableFigure 5, and removed some variablesKD, GSPD,BN, EGR, CWPM, GPR with high correlation and low feature importance same as CART above. The model has accuracy 64%, AUC 66%, precision 63%, and recall 64%

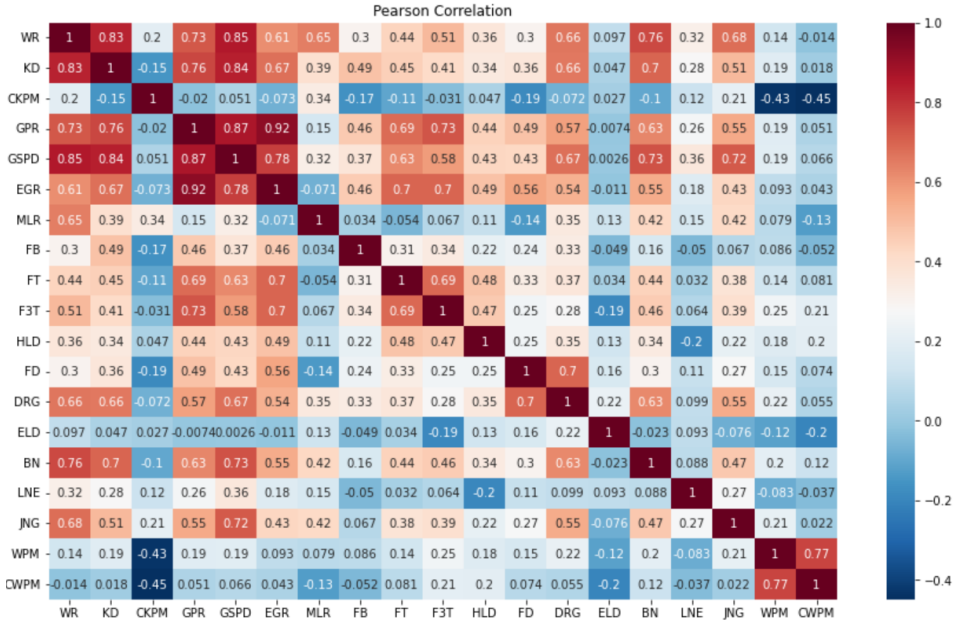

Figure 5: Correlations Diagram for All Variables

The winner of the Random Forest model prediction shown below(Figure 6) was JDG.

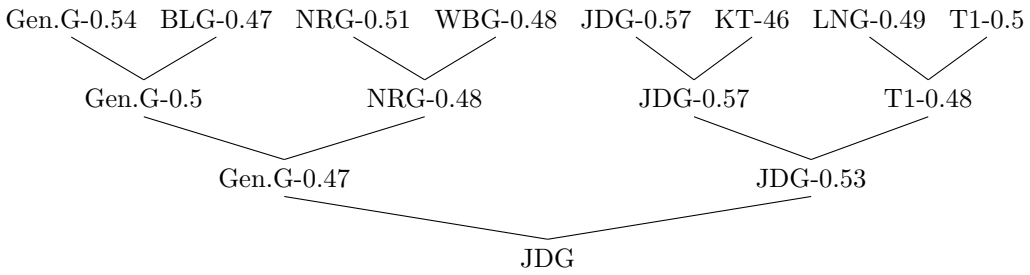

Figure 6: Tree Diagram for Predicting Winner by Random Forest Model

4.4 Boosting

Random Forest is more effective on unstructured data, while XGBoost tends to perform better on structured data, which uses a collection of decision trees and gradient boosting to make predictions. In our fitting, similarly removed the variables with low significance values and obtained a model with an accuracy of 54%, The winner of the XGBoost model prediction shown below(Figure 7) was T1.

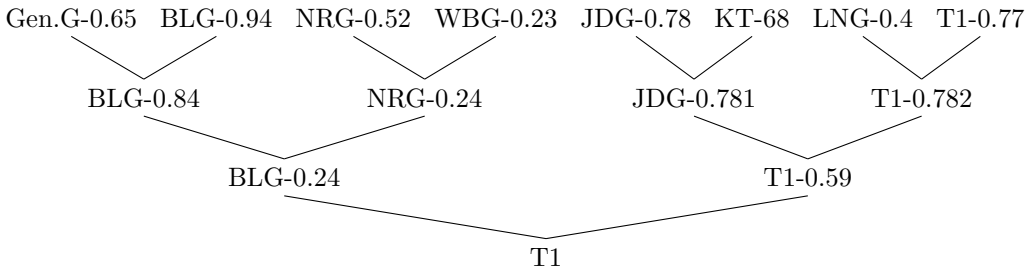

Figure 7: Tree Diagram for Predicting Winner by XGBoost Model

4.5 K-mean

After a simple application of supervised learning, we would like to do some classification of the patterns of the race, where k-mean is applied. after performing the elbow method (Figure 7) and the detection of Silhouette scorek(clusters) = 4 is 0.37 which is the highest, we can determine that 4 is an appropriate number of classifications

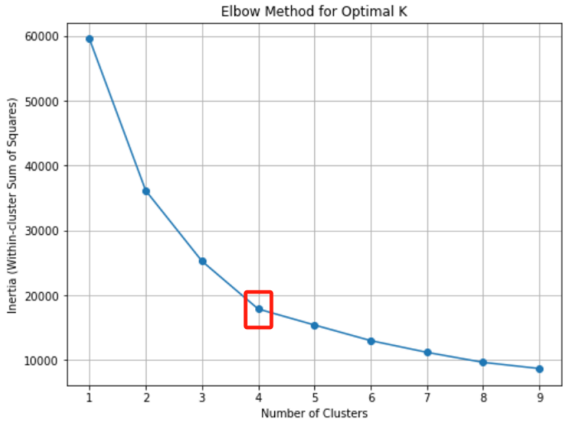

Figure 7: elbow method for finding K for K-mean

After adding K-mean clustering label as a new feature, we then apply the four previous methods to the new dataset and compare the accuracy before and after clustering, and we can see from Table 2 that there are only some small improvements.

	Accuracy	
Model	Before	After
Logistic model	0.621622	0.648649
CART	0.523461	0.576577
Random Forest	0.622222	0.662979
Boosting	0.540541	0.554054

Table 2: The accuracy for each model Before and after applying K-mean label

In the new four-model trial, we predicted only the finals, i.e., WBG and T1. Logistic Regression's model showed T1 winning 85% of the time in blue while WBG won 2% of the time in blue, CART's model showed T1 winning 88% of the time in blue while WBG won 93% of the time in blue, Random Forest's model shows T1 winning 51% of the time in blue vs. 49% for WBG, boosting's model shows T1 winning 70% of the time in blue vs. 30% for WBG.

5 Bootstrapping

In summary, we have a total of 8 models, in order to be able to systematically observe the play of each model, we use the bootstrapping method, the 8 models in different datasets will be trained 1000 times each to observe the accuracy of distribution.

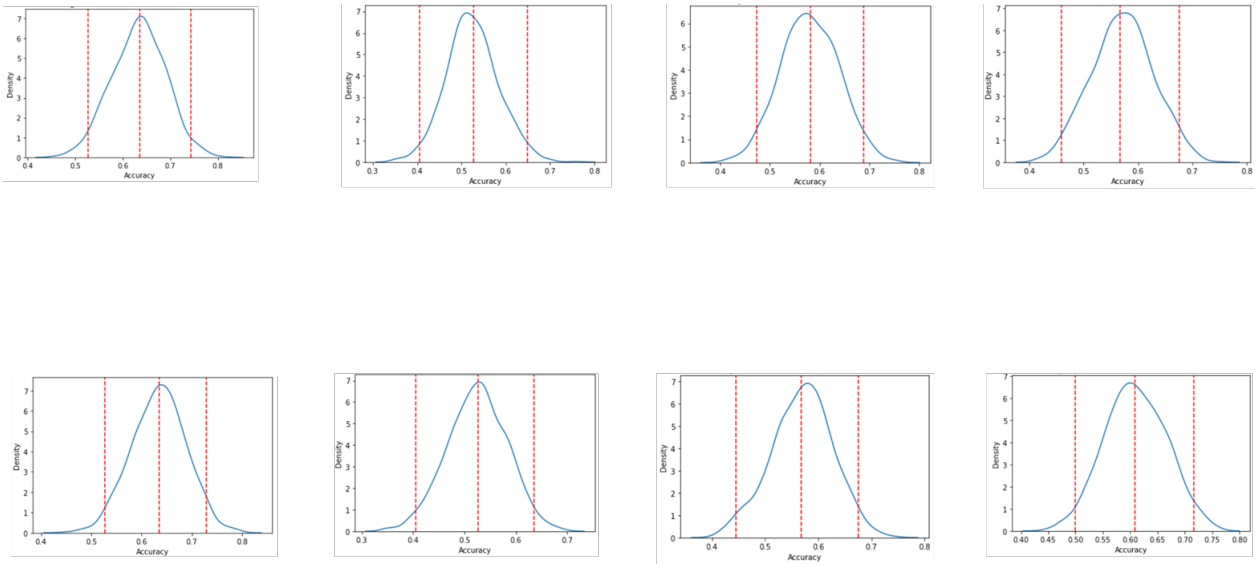

Figure 8: Bootstrapping accuracy distribution with 1000 times

Note: Upper for model fitted with Original dataset and Lower for added-K-mean-feature dataset, the sequence of model from left to right is Logistic Regression, CART, Random Forest and Boosting correspondingly

As can be seen in Figure 8, the addition of the k-mean's feature is not effective in making an improvement in the accuracy of the model, except for boosting, which goes from 55% to 61%. Meanwhile, the graphs of Logistic Regression and random forest show that their standard division is lower, and the overall graph of normal distribution is more clustered, which represents that these two models play more stable.

6 Conclusion

Logistic Regression, CART, Random Forest and Boosting models were used to predict the original data and the data with new labels after k-mean clustering. When comparing the models in bootstrapping, it can be seen that Logistic Regression (64.8%) and Boosting with k-mean label (61%) have better performance.

Logistic Regression had the highest accuracy in predictions without the kmean label, while Boosting successfully predicted the winner to be T1 and succeeded in all but the NRG and WBG match predictions. In the k-mean labeled finals-only prediction, Logistic Regression, CART and Boosting all had T1 winning more than WBG, but given that it was actually T1 3:0 WBG, Random forest's prediction of the winning percentage between the two teams wasn't that different. Overall, Logistic Regression and Boosting played the best in this prediction.

7 Discussion

Interesting fact:

In the four Top 8 predictions using the unadded k-mean label, which is the original collated dataset, we can see that in the WBG and NRG predictions all the models had WBG on the losing side, yet in fact WBG held on to play T1 in the finals, which makes WBG a team that broke the predictions.

Advantage:

Our model has a relatively large number (19) of features, which basically contains all the data that can be measured in the game. The prediction total of 8 metrics of the model, including accuracy, TPR, FPR, AUC, etc. also gives the user a general direction. At the same time, each team can play about 100 games per year, we extracted the data for the average of the year as the prediction of the play during the game is similar to a model, then the prediction of the winning percentage for the model on the model, which is a more complex model.

Improvement:

Some teams have good yearly average stats but don't do well in Worlds because of the level gap between regions, so the difference between regions can be used as a variable for the future. Also, the effect of losing or winning matches in different formats such as best-of-three or best-of-five on a team's performance in the next match, maybe we can apply time-series later on.

8 Reference

[1] “How to Play - League of Legends.” How to Play - League of Legends, www.leagueoflegends.com/en-us/how-to-play/. Accessed 6 Dec. 2023.

[2] “League of Legends.” Wikipedia, Wikimedia Foundation, 26 Nov. 2023, en.wikipedia.org/wiki/League_of_Legends.

[3] Lol Esports, lolesports.com/article/state-of-the-game-lol-esports-in-2023/blt5d3bca31d1b39e0c. Accessed 6 Dec. 2023.

[4] Steph RoehlerSteph, “Why Is the Blue Side Better in League of Legends?” TRN Checkpoint, 11 May 2023, tracker.gg/checkpoint/articles/why-is-the-blue-side-better-in-league-of-legends.

[5] https://oracleselixir.com/

[6] https://lol.fandom.com/wiki/League_of_Legends_Esports_Wiki

Figure 1: https://imgsvc.trackercdn.com/url/size(1280x720),fit(cover),quality(100)/https%3A%2F%2Ftrackercdn.co

Figure 2, 5: see in jupyter notebook: https://github.com/xilin-tian/Winner_Prediction_project/blob/main